The Expanded Racial and Ethnic Codes in the Medicare Data Files: Their Completeness of Coverage and Accuracy

Diane S. Lauderdale, MA, and Jack Goldberg, PhD

ABSTRACT

Objectives. This paper evaluates the new race/ethnicity codes for Asian Americans, Hispanics, and Native Americans that have recently been added to the Medicare enrollment database.

Methods. The race/ethnicity code revisions made by the Health Care Financing Administration are described and evaluated by (1) comparing the numbers of persons identified as Asian Americans, Hispanics, and Native Americans with corresponding population census projections and (2) determining whether Medicare enrollees born in Asian and Hispanic countries are assigned Asian and Hispanic codes.

Results. Among persons 65 years of age and older, approximately 24% of Hispanics, 17% of Native Americans, and 56% of Asian Americans are identifiable by the new codes. From 18% to 29% of enrollees 65 years old or older born in Mexico, Puerto Rico, and Cuba are coded as Hispanic, and from 14% to 73% of enrollees born in nine Asian countries are classified as Asian American. Classification is not random but is related to timing of migration and to country of origin.

Conclusions. Researchers should resist the temptation to base analyses on the revised Health Care Financing Administration race/ethnicity codes, since coverage is incomplete and biased. (*Am J Public Health.* 1996;86:712–716)

Introduction

> *"What's the use of their having names,"* the Gnat said, *"if they won't answer to them?"*
>
> *"No use to* them," *said Alice; "but it's useful to the people that name them, I suppose."*
>
> Lewis Carroll[1]

Prior to July 1994, the Health Care Financing Administration (HCFA) maintained records for Medicare enrollees coded in one of four race categories: White, Black, other, or unknown. Race is one of the few demographic variables available on Medicare records and has been a focus of many of the varied studies based on Medicare data. The structure of analysis in evaluations of hospital mortality; examinations of procedures, outcomes, and use; disease surveillance; and epidemiology has been shaped by, and limited by, these four categories. Typical studies compare Black and White enrollees only, ignoring the "other" group because of its heterogeneity. Recent research includes a comparison of Medicare claims by race based on 32 procedures and tests[2]; an analysis of poststroke survival by medical and demographic factors, including race[3]; and calculation of hip fracture incidence rates by age, sex, and race.[4]

The shortcomings of Black/White/other/unknown race data (i.e., inconsistency with other federal agencies) was apparent to those both within and outside HCFA. A lawsuit was filed in 1993 against the secretary of health and human services and the Department of Health and Human Services alleging violation of Title VI and the Civil Rights Act in the failure to provide racial and ethnic identifiers for monitoring civil rights compliance.[5] In July 1994, HCFA expanded the race code. Three codes were added: Asian, Asian American, and Pacific Islander; Hispanic; and North American Indian and Alaskan Native.

This paper consists of two parts. First, we present a historical summary of race data at HCFA. Second, we evaluate whether the new codes identify the populations that would be expected by a researcher; particular attention is focused on Asian and Hispanic groups. We do not consider whether conflating four race groups and one ethnicity (Hispanic) is either appropriate for public health research or theoretically defensible.[6] Although race and ethnicity are distinct concepts, with many academics eschewing the former altogether,[7-10] we refer to the HCFA data item here as race/ethnicity, reflecting its content. We conclude with a discussion of the present utility of and future prospects for identifying Asian Americans, Native Americans, and Hispanics with HCFA race/ethnicity codes.

History of Race Coding

HCFA enrollment database records are not created by the agency itself when a person enrolls in the Medicare program; rather, data are transferred from either the Social Security Administration or the Railroad Retirement Board. The Railroad Retirement Board records, for persons whose eligibility for Medicare is

The authors are with the Division of Epidemiology-Biostatistics, School of Public Health, University of Illinois–Chicago. Jack Goldberg is also with the Vietnam Era Twin Registry and Center for Cooperative Studies in Health Services, Department of Veterans Affairs Medical Center, Hines, Ill.

Requests for reprints should be sent to Jack Goldberg, PhD, University of Illinois at Chicago, School of Public Health, 2121 W Taylor, SPH-West, Room 534 (M/C 922), Chicago, IL 60612.

This paper was accepted September 26, 1995.

based on railroad employment, do not contain race information and account for many of the 3.5% of enrollment records involving unknown race.[11] The vast majority of enrollment database records come from the master beneficiary record file of the Social Security Administration; Medicare's race information reflects that available from its source file.

Race information is collected by the Social Security Administration on a voluntary basis on the application form for a Social Security number (form SS-5). (The same information is also collected on applications for replacement cards and notifications of changed information, such as a new surname upon marriage.) From the inception of the Social Security program in 1936 until 1980, the SS-5 form included three categories for race: White, Black, and other. Thus, when the master beneficiary record file was established, race was coded as White, Black, other, or unknown. In November 1980, the race choices on the SS-5 (and related forms) were broadened to comply with Office of Management and Budget directive 15, (race and ethnic standards for federal statistics and administrative reporting). The stated objective of the directive was to standardize classifications for federal administrative and statistical purposes.[12] Unstated, but surely also important, was the imperative to collect and report data that could better accommodate the changing racial and ethnic composition of the population.

Directive 15 does not define race or ethnicity, avoiding scientific issues, but it presents rules for classifying persons into five groups, based primarily on geographic origin.[13,14] The directive offers two alternatives for identifying Hispanic ethnicity. The preferred method is to collect information on race and Hispanic ethnicity in separate questions, so that all Hispanic persons also have a stated race. The alternative is to list Hispanic as a choice on a single race/ethnicity question, qualifying both "Black" and "White" with the phrase "not of Hispanic origin." The Social Security Administration chose this second option; the "other" category on the SS-5 form was replaced by the following three categories: Asian, Asian American, or Pacific Islander; Hispanic; and Northern American Indian or Alaskan Native. Surprisingly, the Social Security Administration did not restructure the master beneficiary record from the White/Black/other/unknown format; thus, none of the new information found its way into the enrollment database at HCFA.

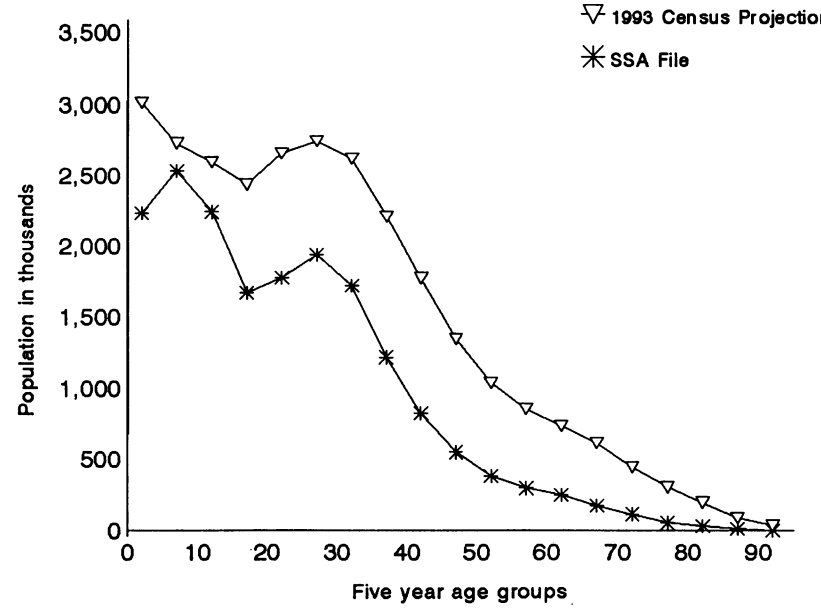

FIGURE 1—Estimated 1993 US Hispanic population, by age group: Bureau of the Census projections and Social Security Administration (SSA) numerical identification file.

The great majority of today's elderly population (65 years of age or older) applied for Social Security numbers well before 1980. Many have had no occasion since 1980 to request a replacement card or file a change of information. For this group of the elderly population, the Social Security Administration has only the race information collected on the earlier form. Conversely, some elderly persons (e.g., those who never entered the labor force or who immigrated recently to the United States) have applied for Social Security numbers since 1980. While these persons used the form with five categories, the categories were collapsed to the White/Black/other format in the master beneficiary record.

Thus, in HCFA's enrollment database, which does not indicate when race information was collected, one effect of directive 15 is the muddling of the meaning of White, Black, and other. In particular, before 1980, Hispanics who did not think of their Hispanic identity as racial would check White or Black; since 1980, Hispanics have been instructed not to check White or Black, and these individuals constitute a major component of the "other" category. This compounds inconsistencies arising from shifts in ethnic self-identity, such as an increased tendency for American Indians to identify themselves as such rather than as White.[15,16]

HCFA developed a two-step strategy to improve its race data. The first step, implemented in July of 1994, was to transfer race/ethnicity data directly from the SS-5 file at the Social Security Administration—the numerical identification file—bypassing the master beneficiary record for this one data item. The second step of HCFA's strategy called for a mailing to all persons whose race remained classified as other or unknown; the mailing requested new race/ethnicity information.

The numerical identification file does not simply contain a single SS-5 form for each person. SS-5 forms were removed from the file prior to its conversion to a machine-readable format in the 1970s when a person filed for a claim. Replacement records did not include race. Thus, SS-5 race data are unavailable for about one in five elderly persons; conversely, persons who have applied for replacement or new cards may have more than one record, with conflicting race/ethnicity data. However, obtaining this information from the numerical identification file does avoid two problems with the master beneficiary record race item. First, race in the master beneficiary record always applies to the person whose work record

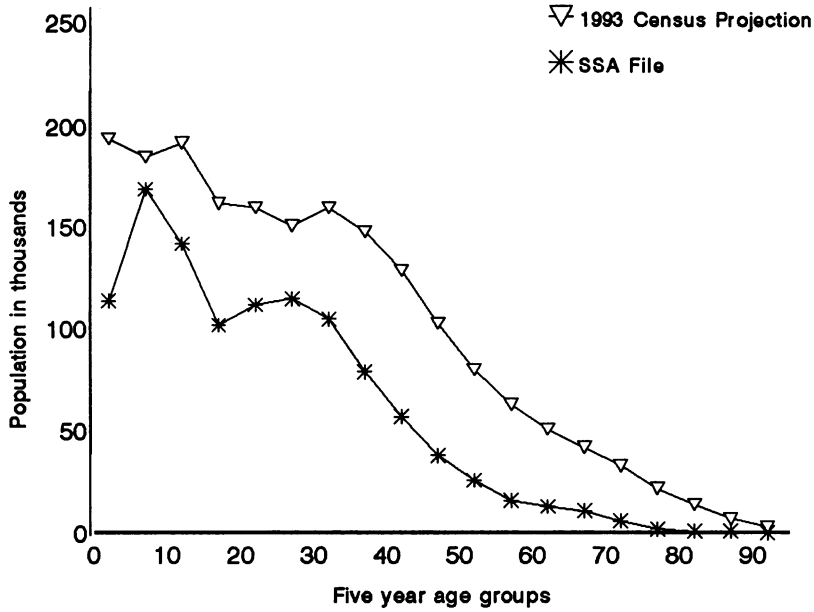

FIGURE 2—Estimated 1993 US Native American population, by age group: Bureau of the Census projections and Social Security Administration (SSA) numerical identification file.

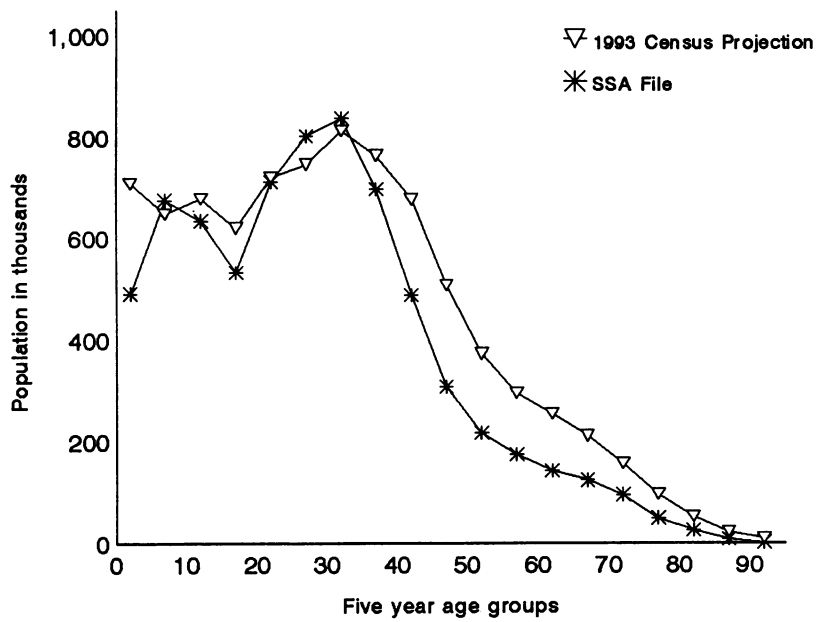

FIGURE 3—Estimated 1993 US Asian American population, by age group: Bureau of the Census projections and Social Security Administration (SSA) numerical identification file.

forms the basis of entitlement to Social Security benefits or Medicare eligibility. Thus, the race listed on the master beneficiary record of a wife eligible as the auxiliary of her husband is that of the husband. Second, a recent defect in data transfer from the SS-5 form to the master beneficiary record resulted in the incorrect coding of persons as other or unknown who actually are White or Black (written communication, B. Kestenbaum, Social Security Administration, September 1994).

The implementation of the first step of HCFA's strategy resulted in changes of race/ethnicity coding for more than 2.5 million enrollees. Most changes were corrections. Only three quarters of a million enrollees were reclassified into one of the three new categories: about 500 000 were reclassified as Hispanic, about 200 000 were reclassified as Asian American, and about 40 000 were reclassified as Native American (written communication, HCFA, July 1994). About three quarters of those now classified as Asian American and slightly more than half of those now classified as Native American were previously classified as other. In contrast, about two thirds of those now classified as Hispanic were previously classified as White.

Materials and Methods

The first part of the analysis assesses the completeness of coverage for HCFA race/ethnicity information by using a 1-in-100 sample of numerical identification file records. The number of residents in the 50 states and Puerto Rico classified as Asian American, Hispanic, or Native American in the numerical identification file is compared with the corresponding projected counts based on the 1990 census of population.[17] The comparison is made over the entire age range, although our primary interest is in the older population. To adjust for multiple records in the numerical identification file, it was necessary to constrain the totals by age group to agree with census totals by age group while maintaining the relative distributions by race within age groups.

The second part of the analysis examines the accuracy of HCFA's race/ethnicity classification by using a 1-in-100 sample of numerical identification file records linked with the master beneficiary record. The master beneficiary record indicates Medicare enrollment, while the numerical identification file includes country of birth (unavailable at HCFA). The focus is on the distribution of race codes by country of birth for persons enrolled in Medicare (not on the basis of railroad employment) at the end of 1993. For Medicare enrollees from each of several Hispanic and Asian birthplaces, the proportions identified by each race/ethnicity code in the numerical identification file

are calculated. These are the codes now available from the enrollment database at HCFA. The tabulations test the expectations that almost all persons born in Hispanic countries will be classified as Hispanic and that almost all persons born in Asian countries will be classified as Asian. Observed variations in these proportions for Asian countries are explored in terms of the timing of migration from each country relative to 1980.

Results

Completeness of Coverage

The numbers, by 5-year age groups, of Hispanics, Native Americans, and Asians living in the United States and Puerto Rico and identified in the numerical identification file in September 1993 are compared with corresponding Bureau of the Census projections in Figures 1, 2, and 3, respectively. For all three ethnic groups, coverage was more extensive for the younger ages and less so for the older ages. Focusing on the age groups from 65 to 94 years, numerical identification file coverage relative to the census was 24% for Hispanics, 17% for Native Americans, and 56% for Asians. The youngest age group (0 to 4 years) is an anomaly for all groups because the new Social Security Administration enumeration at birth initiative, which assigns Social Security numbers to newborns, fails to collect race/ethnicity information.[18]

For both Hispanics and Native Americans, coverage decreased monotonically with increasing age. This differed from the pattern of coverage for Asians, Asian Americans, and Pacific Islanders, which was more complete at all ages and did not display so consistently the inverse relationship between age and coverage.

Accuracy of Numerical Identification File Race Codes

The distributions of race/ethnicity codes for elderly persons enrolled in Medicare at the end of 1993 and born in three selected Hispanic and nine Asian countries are presented in Tables 1 and 2. (Country of birth and race are missing for persons whose SS-5 form had been removed to use in the adjudication of a claim; therefore, proportions here with unknown race are much lower than the 20% previously noted.)

The majority of elderly Medicare enrollees born in Cuba, Mexico, and Puerto Rico were classified in the numerical identification file not as Hispanic but as White. The highest Hispanic proportion (.29) was that for enrollees born in Mexico. This finding can be immediately extended to all elderly Cubans in the United States but not to all Mexicans, since 98% of elderly persons with Cuban ancestry, but only 39% with Mexican ancestry, were foreign born.[19] Because Puerto Rico participates in the Social Security and Medicare programs, most persons in the numerical identification file born in Puerto Rico are still residents of Puerto Rico.

Among the nine Asian countries of birth, the proportions classified as Asian American varied widely, from .14 for Japan to .73 for Cambodia. For all countries, most persons not classified as Asian were identified as other; the proportions classified as White were low, except for India. Historically, the race of persons from India has been ambiguous; for example, the 1970 census included Asian Indians in the White category, while Hindu was a separate race in the 1930 and 1940 censuses.[20,21] For all of these countries except Japan, the results obtained for immigrants now 65 years of age or older may be generalized to essentially all elderly persons with that ancestry, because the overwhelming majority were foreign born.[22]

The variation from country to country in the proportion identified as Asian can be explained by the timing of migration, that is, whether persons immigrated before or after the Asian category was added to the SS-5 form in 1980. For the Southeast Asian countries, for example, peak years of emigration followed 1980, key years being 1980 through 1983 for Laos, 1981 through 1987 for Cambodia (Kampuchea), and 1978 through the decade of the 1980s for Vietnam.[20] Confirming this relationship was a (weighted) correlation coefficient of .90 between the proportions of elderly Medicare enrollees classified as Asian in the numerical identification file from each country and the proportions of emigrants, according to the census, who were 62 years old or older in 1990 and who had immigrated since 1980.

TABLE 1—Race/Ethnicity Coding in the Social Security Administration Numerical Identification File for a 1% Sample of Medicare Enrollees Born in Mexico, Puerto Rico, and Cuba

Country of Birth	Total No.	Hispanic, %	White, %	Black, %	Other, %	Unknown, %	All Races, %
Cuba	1640	18	74	3	2	3	100
Mexico	2084	29	58	0	8	4	100
Puerto Rico	3661	25	55	4	6	9	100

Note. Asian and Native American codes are not shown. Percentages may not add to 100 because of rounding.

TABLE 2—Race/Ethnicity Coding in the Social Security Administration Numerical Identification File for a 1% Sample of Medicare Enrollees Born in Asian Countries

Country of Birth	Total No.	Asian, %	Other, %	White, %	Unknown, %	All Races, %
Cambodia	40	73	28	0	0	100
China[a]	963	33	57	6	0	100
India	160	48	28	18	6	100
Japan	194	14	74	7	3	100
South Korea	266	52	44	2	3	100
Laos	58	52	31	7	10	100
Pakistan	26	50	38	3	3	100
Philippines	1002	36	53	5	5	100
Vietnam	36	72	17	6	0	100

Note. Black, Hispanic, and Native American codes are not shown. Percentages may not add to 100 because of rounding.
[a]Includes Taiwan and Hong Kong.

Discussion

Race and ethnicity, along with sex and age, are routinely used as categories in descriptive studies and surveillance activities and as covariates in analytic research; they have repeatedly been found to be associated with morbidity and mortality. *Healthy People 2000* stresses the need for race and ethnicity data in public health.[23]

Since 1980, the Social Security Administration has been collecting race/ethnicity information that promises to identify Hispanics, Asians, and Native Americans, and now HCFA is drawing on these data. With respect to the elderly population, however, the expanded classification scheme has not been in use long enough to substantially identify the target groups, as is evident from a comparison with census population data. Furthermore, the Black, White, and "other" categories have different meanings for the pre-1980 and post-1980 records.

One weakness in the use of the numerical identification file to assess the new codes is the problem of missing SS-5 forms. Because most forms are missing as a result of claim adjudication before the file was transferred to a machine-readable format, the missing forms had the earlier three-category race item. Thus, no additional persons with the new codes would have been identified if the forms had not been removed. Consequently, our evaluation of accuracy by birthplace probably overestimates the percentages identifiable by the new codes for persons born in Hispanic and Asian countries since missing foreign-born individuals would all have Black/White/other/unknown codes.

The incompleteness of the new race/ethnicity codes is much greater for Hispanics and Native Americans than for Asians. For Asians, however, identification by the Asian American category is not random but rather is related to the timing of migration and hence to the country of origin.

In the second step of its strategy to improve the identification of racial and ethnic groups, HCFA is undertaking a mail census of enrollees still coded as other (or with race unknown). Our results imply that this initiative offers promise for significant improvement in the identification of persons born in Asian countries but not of those with Hispanic birthplaces, most of whom are presently classified as White.

Our findings have ramifications for public health researchers using HCFA data to study racial and ethnic minorities (other than Blacks). If numerators are derived from Medicare files and denominators are derived from another source, such as the Bureau of the Census, rates will be badly understated. If both numerator and denominator are drawn from Medicare files, rates will be biased: recent immigrants will be overrepresented in the data, and earlier immigrants and native-born individuals will be underrepresented. Needless to say, recent immigrants are likely to have different histories and health behaviors than those who have resided in the United States for longer periods. For example, the earlier wave of Cuban immigrants included many wealthy, professional persons, while more recent Cuban immigrants are generally from less affluent backgrounds. Among immigrants from Asia, there will be a country effect: immigrants from Southeast Asian countries, many of whom are war refugees, will be disproportionally represented relative to Chinese, Filipinos, and, especially, Japanese. Similarly, case–control studies with controls drawn by means of these codes and cases selected from a clinical setting may also involve biased results.

Researchers will need to resist the temptation to routinely generate disease, hospitalization, procedure, and mortality rates for elderly Hispanics, and, pending the results of HCFA's mailing to persons classified as other, they will need to exercise similar restraint for elderly Asians. □

Acknowledgments

This study was supported by National Institute on Aging grants T32 AG00231 and P20 AG12042.

We thank Bert Kestenbaum, Office of the Actuary, Social Security Administration, for technical assistance with record linkage and for significant substantive and editorial comments.

References

1. Carroll L. *The Annotated Alice: Alice's Adventures in Wonderland & Through the Looking Glass.* New York, NY: Clarkson N. Potter, Inc.;1960:222.
2. Escarce JJ, Epstein KR, Colby DC, Schwartz JS. Racial differences in the elderly's use of medical procedures and diagnostic tests. *Am J Public Health.* 1993; 83:948–954.
3. May DS, Casper ML, Croft JB, Giles WH. Trends in survival after stroke among Medicare beneficiaries. *Stroke.* 1994;25: 1617–1622.
4. Jacobsen SJ, Goldberg J, Miles TP, Brody JA, Stiers W, Rimm AA. Hip fracture incidence among the old and very old: a population-based study of 745,435 cases. *Am J Public Health.* 1990;80:1374–1380.
5. Williams DR, Lavizzo-Mourey R, Warren RC. The concept of race and health status in America. *Public Health Rep.* 1994;109: 26–41.
6. Senior PA, Bhopal R. Ethnicity as a variable in epidemiologic research. *BMJ.* 1994;309:327–330.
7. Gould SJ. Why we should not name races: a biological view. In: Gould SJ, ed. *Ever since Darwin.* New York, NY: WW Norton & Co; 1977:231–236.
8. Cooper RS. A case study in the use of race and ethnicity in public health surveillance. *Public Health Rep.* 1994;109:46–52.
9. Cruikshank JK, Beevers DG. Preface. In: Cruikshank JK, Beevers DG, eds. *Ethnic Factors in Health and Disease.* Oxford, England: Butterworth-Heinemann Ltd; 1989:vii–ix.
10. Cavalli-Sforza LL, Menozzi P, Piazza A. *The History and Geography of Human Genes.* Princeton, NJ: Princeton University Press; 1994;19–20.
11. Hatten J. Medicare's common denominator: the covered population. *Health Care Financing Rev.* 1980;2:53–63.
12. Office of Management and Budget. Directive no. 15: race and ethnic standards for federal statistics and administrative reporting. In: *Statistical Policy Handbook.* Washington, DC: Office of Federal Statistical Policy and Standards, US Dept of Commerce; 1978:37–38.
13. Hahn RA. The state of federal health statistics on racial and ethnic groups. *JAMA.* 1992;267:268–271.
14. Hahn RA, Stroup DF. Race and ethnicity in public health surveillance: criteria for the scientific use of social categories. *Public Health Rep.* 1994;109:7–15.
15. Snipp CM. Who are American Indians? Some observations about the perils and pitfalls of data for race and ethnicity. *Popul Res Policy Rev.* 1986;5:237–252.
16. Passel JS, Berman P. Quality of 1980 census data for American Indians. *Soc Biol.* 1986;33:163–182.
17. *Population Projections of the United States by Age, Sex, Race, and Hispanic Origin: 1992–2050.* Washington, DC: US Bureau of the Census; 1992.
18. Jabine TB. Procedures for restricted data access. *J Off Statistics.* 1993;9:537–589.
19. *1990 Census of Population: Persons of Hispanic Origin in the United States.* Washington, DC: US Bureau of the Census; 1993. Dept of Commerce publication CP-3-3.
20. Barringer HR, Gardner RW, Levin MJ. *Asians and Pacific Islanders in the United States (The Population of the United States in the 1980s).* New York, NY: Russell Sage Foundation; 1993.
21. Lee SM. Racial classification in the US census: 1890–1990. *Ethnic Racial Stud.* 1993;16:75–94.
22. *1990 Census of Population: Asians and Pacific Islanders in the United States.* Washington, DC: US Bureau of the Census; 1993. Dept of Commerce publication CP-3-5.
23. *Healthy People 2000: National Health Promotion and Disease Prevention Objectives.* Washington, DC: US Dept of Health and Human Services; September 1990. DHHS publication PHS 91-50212.